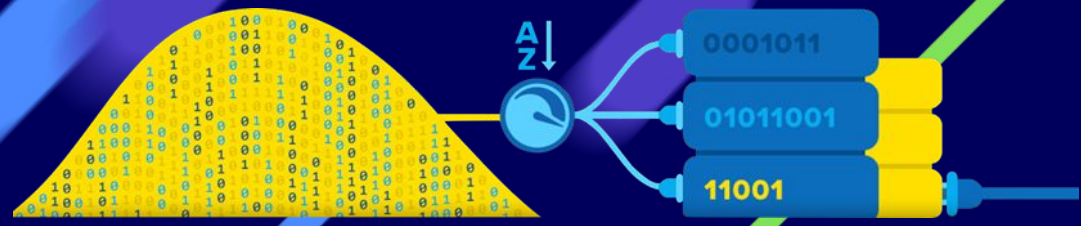
Обзор программного обеспечения Trillium

Бен Данмор

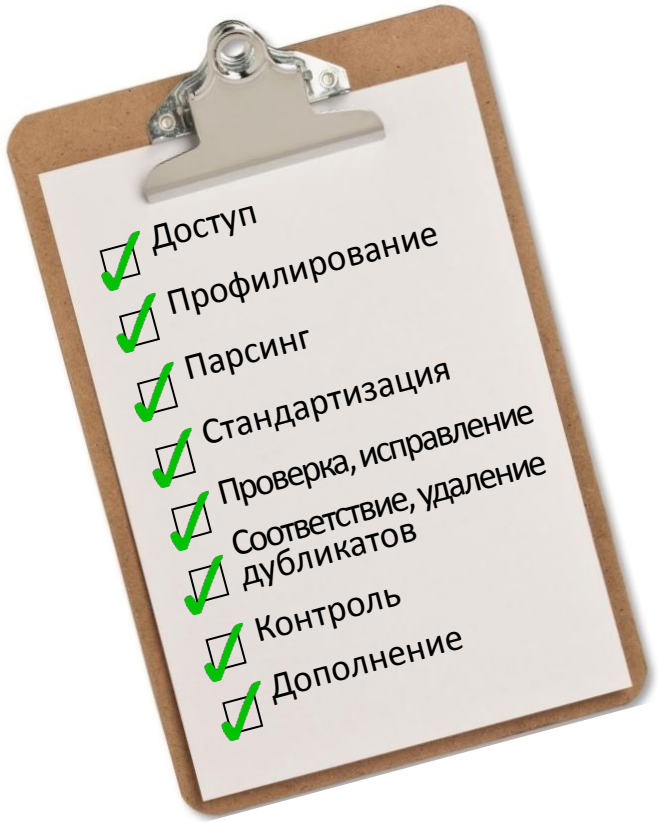
Старший консультант по предпродажам

16 мая 2018 г.



Чем мы занимаемся

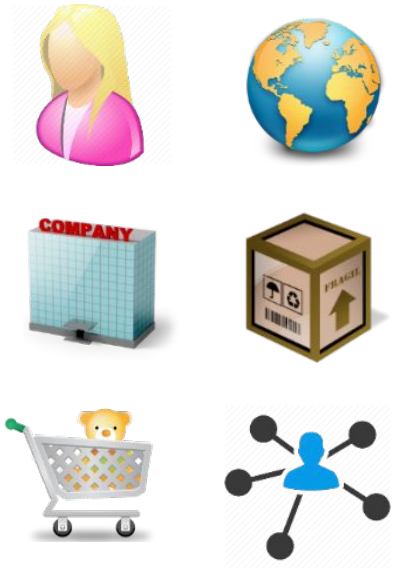
Мы продаем программное обеспечение и услуги, **способствующие повышению качества данных** и **их единообразию.**



Совместимость со всеми типами данных

Программные решения Trillium могут обрабатывать данные, независимо от их…

типа



формата

более 50 кодовых страниц

источника

территории использования

более 200 территорий

объема

Высокая масштабируемость:
зафиксировано свыше
100 миллионов записей в час!

Архитектура

Центр изучения

Центр управления «толстыми» клиентами

Создание отчетов, управление

Пользовательский доступ

TRILLIUM SOFTWARE

Связь источников данных

Пакетная обработка

«Большие данные»

Доступ в реальном

Как это работает? Наша методология

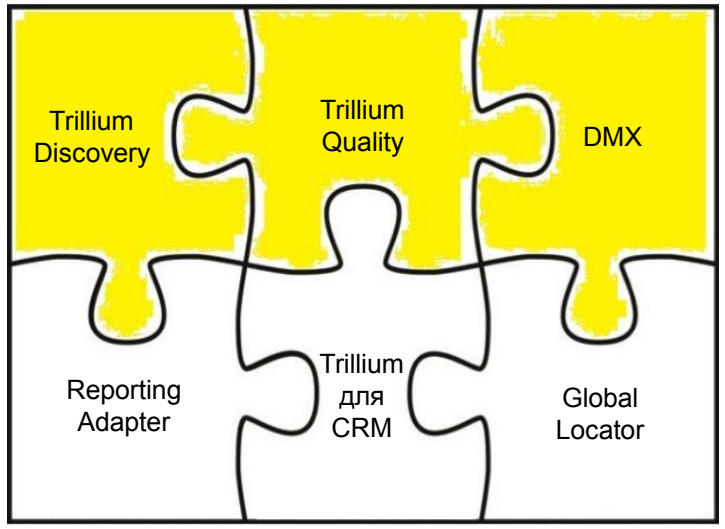
1) ОБНАРУЖЕНИЕ
Анализ и проверка данных
Выявление проблем с качеством
Составление норм и правил
Количественная оценка и ранжирование проблем

2) РАЗРАБОТКА
Схемы обеспечения качества данных
Преобразование, слияние
Очистка, дополнение
Соответствие

4) УПРАВЛЕНИЕ
Анализ целевых показателей, отчетов, оценочных листов, графиков
Анализ и улучшение предоставляемых услуг обеспечения качества данных
Управление серверами обеспечения качества данных
Сопровождение клиентов

3) РАЗВЕРТЫВАНИЕ
Подготовка конфигурации, работающей в реальном времени
Проверка параметров времени исполнения
Проверка процессов
Выпуск услуг обеспечения качества данных

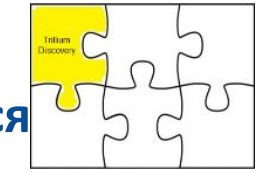
6 основных продуктов



- **Trillium Discovery** Удобное средство профилирования данных
- **Trillium Quality** Разработка и развертывание процессов очистки и сопоставления
- **DMX и DMX-h** Комплексное средство извлечения, преобразования и загрузки данных (ETL), в том числе доступ к мейнфрейму с помощью технологии Hadoop
- **Reporting Adapter** Простое подключение к средствам бизнес-аналитики и управления сторонних поставщиков
- **Trillium для CRM** Подключаемые модули обеспечения качества данных для MS Dynamics и SAP
- **Global Locator** Определение адресов по всему миру и их проверка в реальном времени или с помощью пакетной обработки

Trillium Discovery

Основные проблемы, с которыми помогает справиться Discovery



выборка и охват данных — какое количество данных содержит каждое поле?

расположение данных — в правильном ли поле расположены данные?

структура, соответствие шаблону — соответству-ют ли данные формату или ограничению по

целостность данных — соответствуют ли данные правилам, установленным для бизнеса?

уникальность — уникальны ли основные и внешние ключи?

не **«подделаны»** ли данные — не внес ли кто-нибудь ложные данные, чтобы обойти обязательные

повторы — какое количество записей повторяется, какие общие

опечатки — много ли орфографических ошибок содержат данные?

ненормативная лексика — не использовал ли недовольный сотрудник грубые слова, чтобы умышленно повредить ваши данные?

фирменный стиль — соблюдены ли все правила или информация оформлена произвольно?

помогает узнать больше о состоянии своих данных, а именно — **«узнать неизвестное»**

Trillium Discovery: примеры метаданных

Name	Ref	Values	Value Dist %	Patterns	Masks	Min	Max	Min Len	Max Len	Null Count	Null Dist %	Space Count	Space Dist %	Metaphones	Soundexes
Customerid	1	4949	99.839	66	66	001D000000eiaTYIAY	0012000000040436AAA	10	18	0	0.000	0	0	1905	627
Customersource	2	5	0.101	1	1	ACCT	SFDC	4	4	0	0.000	0	0	5	5
Name1	3	3394	68.469	2313	2300	0	龙口泰进机械有限公司	1	108	0	0.000	0	0	2585	875
Name2	4	538	10.853	426	427	2/F, Portion B of 4/F, 5/F & 6/F	Zone A	3	40	4103	82.772	0	0	482	314
Name3	5	381	7.686	313	313	1B, Maruthandapalli	Zhongshan Viasystems Printed Circuit	3	40	4505	90.882	0	0	299	236
Name4	6	1190	24.006	1085	1084	1 North Section Digital Road	Üniversiteler Mah.İhsan Doğramacı Bul.	2	40	3394	68.469	0	0	1054	655
Street	7	3295	66.472	2663	2631	0	영등포구 여의대로 24, FKI타워	1	209	1053	21.243	0	0	2672	1225
District	8	1196	24.127	760	763	1 Huanghuai Road, Futian Free Trade Zone	ıs Hanı No:25 Kat:4 Ustbostancı	3	40	3452	69.639	0	0	1114	661
City	9	1259	25.398	425	403	0	진천군	1	40	780	15.735	0	0	895	579
Postcode	10	1565	31.572	40	42	0	102-0083	1	19	1252	25.257	0	0	18	18
Region	11	317	6.395	91	77	0	대한민국	1	21	1848	37.281	439	8.856	164	140
Country	12	42	0.847	21	19	00	这中国	2	24	279	5.628	0	0	24	23

Исходные поля

Уникальность

Количество «форматов» данных

Объем и длина каждого атрибута

Количество пустых значений

Записи с пробелами

Использование фонетики для выявления созвучных атрибутов

Эти и МНОГИЕ другие данные предоставляются в **исходном состоянии**!

Trillium Discovery: но все зависит от контекста

Name	Ref	Values	Value Dist %	Patterns	Masks	Min	Max	Min Len	Max Len	Null Count	Null Dist %	Space Count	Space Dist %	Metaphones	Soundexes
Customerid	1	4949	99.839	66	66	001D000000eiaTYIAY	0012000000040436AAA	10	18	0	0.000	0	0	1905	627
Customersource	2	5	0.101	1	1	ACCT	SFDC	4	4	0	0.000	0	0	5	5
Name1	3	3394	68.469	2313	2300	0	龙口泰进机械有限公司	1	108	0	0.000	0	0	2585	875
Name2	4	538	10.853	426	427	2/F, Portion B of 4/F, 5/F & 6/F	Zone A	3	40	4103	82.772	0	0	482	314
Name3	5	381	7.686	313	313	1B, Maruthandapalli	Zhongshan Viasystems Printed Circuit	3	40	4505	90.882	0	0	299	236
Name4	6	1190	24.006	1085	1084	1 North Section Digital Road	Üniversiteler Mah.İhsan Doğramacı Bul.	2	40	3394	68.469	0	0	1054	655
Street	7	3295	66.472	2663	2631	0	영등포구 여의대로 24, FKI타워	1	209	1053	21.243	0	0	2672	1225
District	8	1196	24.127	760	763	1 Huanghuai Road, Futian Free Trade Zone	ıs Hanı No:25 Kat:4 Ustbostancı	3	40	3452	69.639	0	0	1114	661
City	9	1259	25.398	425	403	0	진천군	1	40	780	15.735	0	0	895	579
Postcode	10	1565	31.572	40	42	0	102-0083	1	19	1252	25.257	0	0	18	18
Region	11	317	6.395	91	77	0	대한민국	1	21	1848	37.281	439	8.856	164	140
Country	12	42	0.847	21	19	00	这中国	2	24	279	5.628	0	0	24	23

Здесь действительно должно быть 100 %!

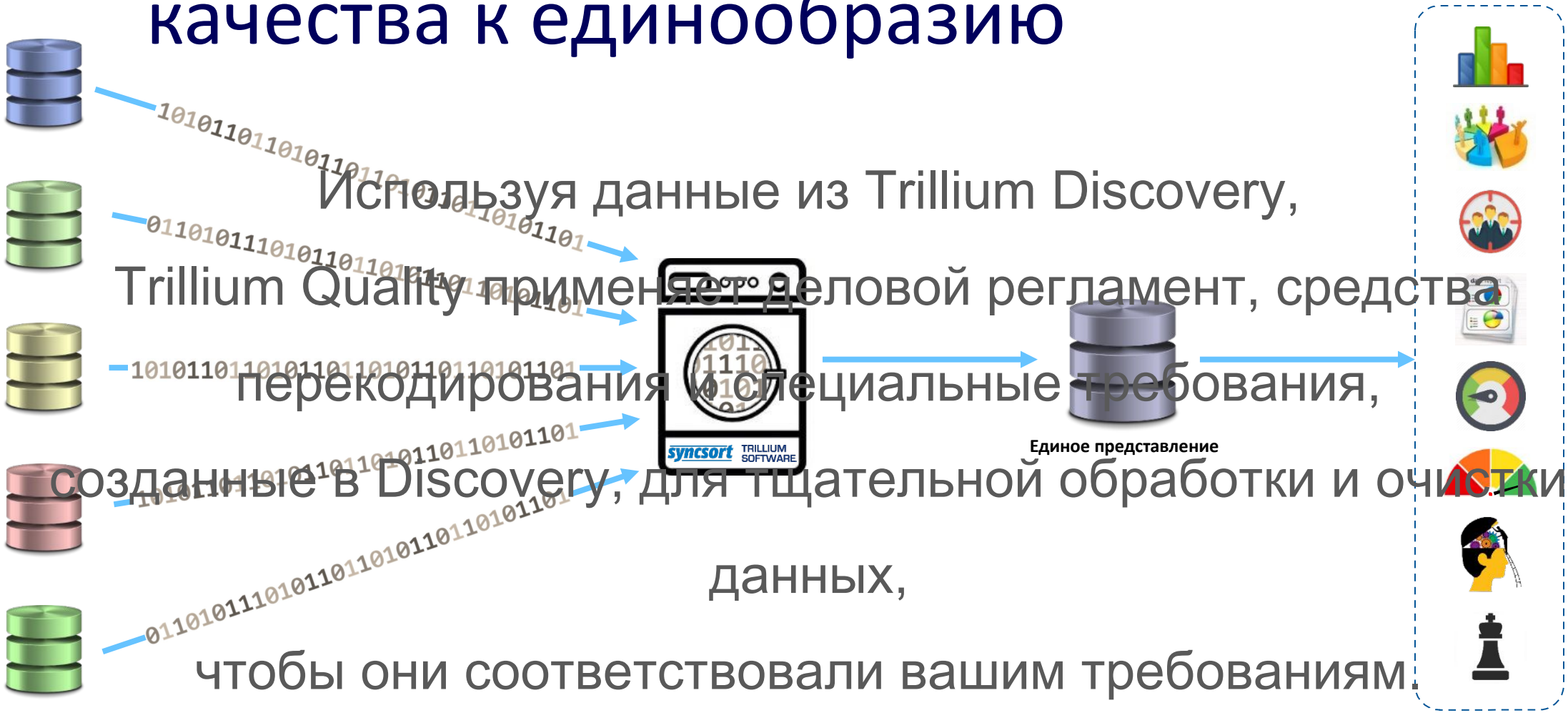
Но никак не в этих строках

Значения NULL допустимы в этих строках

Но их не должно быть здесь

Давайте посмотрим на это приложение в действии...

Trillium Quality – Ваш проект приведения качества к единообразию



Используя данные из Trillium Discovery,

Trillium Quality применяет деловой регламент, средства перекодирования и специальные требования,

созданные в Discovery, для тщательной обработки и очистки данных,

чтобы они соответствовали вашим требованиям.

Trillium Quality: 6 ключевых шагов для обеспечения единообразия

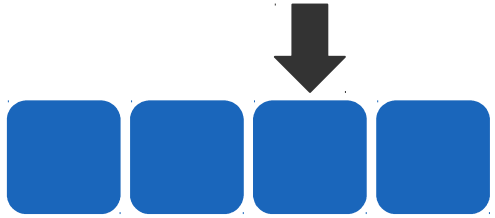
Консолидация

Стандартизация

Парсинг

Очистка, исправление, проверка

Сопоставление, связывание, устранение дубликатов



Согласование и сохранение

1. Консолидация

- Поиск, сбор и **объединение** разных массивов данных в едином централизованном хранилище.
- Данные **не** должны быть одинаковыми: они могут быть из разных источников, иметь разные форматы, длину, названия полей и структуру.

Массив данных 1

Обра-щение	Имя:	Фамилия:	№ дома	Адр. 1	Адр. 2	Адр. 3	Индекс	№ телефона
Mr	Robert	Smith	3	Davy Drive	Rotherham		S66 7EN	01189 407 600

Объединенные данные

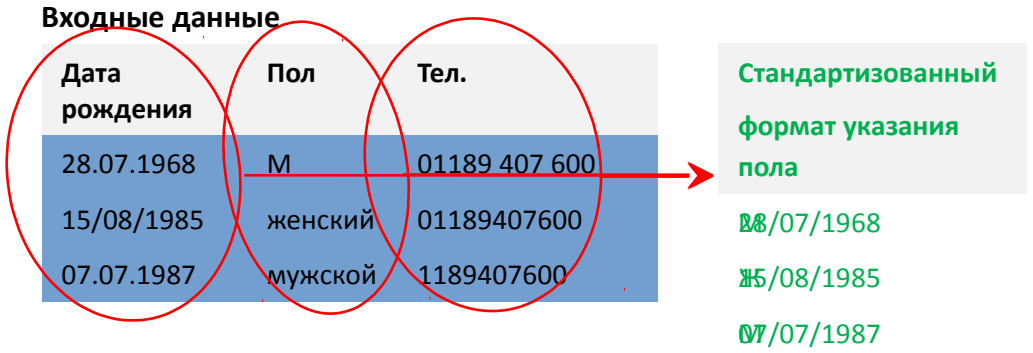
Источник	Обращение	Имя:	Фамилия:	№ дома	Адр. 1	Адр. 2	Индекс	№ телефона
Массив данных 1	Mr	Robert	Smith	3	Davy Drive	Rotherham	S66 7EN	01189 407 600
Массив данных 2		Bob Smith DR			3 Davy Dr		S66 7EN	

Массив данных 2

Полное имя	Адрес 1	Адрес 2	Адрес 3	Почтовый индекс	Телефон
Bob Smith DR	3 Davy Dr			S66 7EN	

2. Стандартизация

- **Стандартизация** данных
- Стандартизованные данные не обязательно заполняют окончательные поля, а скорее помогают при дальнейшем сопоставлении и устранении дубликатов.
- Окончательные поля можно изменять согласно нуждам клиента.



Результат

Исходные данные			Стандартизованные данные		
Дата рождения	Пол	Тел.	Дата рождения	Пол	Тел.
28.07.1968	M	01189 407 600	28/07/1968	M	01189 407 600
15/08/1985	женский	01189407600	15/08/1985	Ж	01189 407 600
07.07.1987	мужской	1189407600	07/07/1987	M	01189 407 600

3. Парсинг

- **Разбор** данных на **составные части**:
  - например, выражение «Dr Bob Smith» разделяется на компоненты «Dr», «Bob», «Smith» в соответствующих отдельных полях.

- Приложение Trillium Quality оснащено умным механизмом, который выполняет стандартизацию данных в контексте:
  - он различает выражения «Bob Smith Dr» и «3 Davy Dr», хотя в первом «Dr» — это ученая степень доктора, а во втором — сокращение от «Drive» (подъездная дорога) в поле для адреса.

Входные данные

Имя:	Адр. 1
Bob Smith DR	3 Davy Dr

Данные после парсинга

Имя:	Адр. 1	Звание	Оригинальное имя	Стандартизованное имя	Фамилия:	Н-р дома	Название улицы	Тип улицы
Bob Smith DR	3 Davy Dr	DR	BOB	ROBERT	SMITH	3	DAVY	DR

4. Очистка, исправление, проверка

- Прогон данных через отраслевые и корпоративные файлы соответствия для **исправления и проверки** позволяет:
  - заполнить пропущенные поля;
  - переместить данные из неправильных полей в правильные;
  - исправить ошибки;
  - дополнить сведения о почтовом индексе в зависимости от зон и районов;
  - использовать правильные поля для указания округов (при необходимости);
  - исправить ошибки регистра: почтовые индексы, названия городов, написанные со строчных букв и т. д.);
  - применить фирменный стиль (например, определить единое написание таких аббревиатур, как Ltd, LLP и т. д.);
  - использовать правильные коды стран.

Входные данные

№ дома	Адр. 1	Адр. 2	Адр. 4	Адр. 5	Индекс
	3 Davy Dr				S66 7EN

Результат (адреса, сверенные с базой данных UK Postal Address File)

Номер дома	Название улицы	Тип улицы	Община	Город	Почтовый индекс
3	Davy	Drive	MALTBY	ROTHERHAM	S66 7EN

5. Сопоставление, связывание, устранение дубликатов

- Поиск потенциально **повторяющихся** записей и перемещение их в специальные **кластеры**.
  - Записи могут совпадать на трех уровнях: индивидуальном, бытовом и корпоративном.
  - Сопоставление не ограничивается количеством и типом полей: могут использоваться как единичные, так и многокомпонентные сопоставления (если A = B и B = C, то A = C).
  - Сопоставлять можно как исходные, так и стандартизованные поля.
  - Можно определять соответствие содержимого, синтаксиса, структур и даже акустических характеристик поля.
  - Пользователи имеют неограниченный контроль параметров сопоставления, выявления подозрительных объектов или ошибок.

Пример кластера

Источник	Имя	Фамилия	Дата рождения	Условие сопоставления
CRM	ROBERT	SMITH	23/05/1988	240
Гарантия	ROBERT	SMITH	23/05/1988	240
Центр обслуживания	ROBERT	SMITH	03/05/1988	242
Маркетинг	ROBERT	SMITH	23/05/1988	240
Финансы	B	SMITH	23/05/1988	241

Условия сопоставления

Категория	Идентификатор шаблона	Имя:	Фамилия:	Дата рождения
P	240	A	A	A
S	241	B	A	A
S	242	A	A	B

6. Согласование и сохранение

- Записи в кластере помогают привести данные в соответствие.
  - Это обеспечивает **единообразие данных во всех записях** и позволяет не оставить данные какой-либо из записей в несоответствующем поле.
- Пользователи могут определить **«эталонную запись»**, используя:
  - последнюю запись;
  - определенный источник данных;
  - ссылочный номер.

Исходные данные

Источник	Имя	Фамилия	Дата рождения	Телефон	Пол	Дата последнего обновления
CRM	Robert	Smith	23/05/1988	01189 407 600	M	21/02/2013
Гарантия	Bob Smith DR		23/05/1988		МУЖСКОЙ	05/01/2018
Центр обслуживания	ROB	SMITH	03.05.1988	01189407600	M	13/07/2012
Маркетинг	Dr Bob Smith		23.05.1988		м	02/05/2014
Финансы	Dr B. Smith		23051988	1189407600	мужской	27/05/2011

Результат стандартизации и унификации

Источник	Имя	Фамилия	Дата рождения	Телефон	Пол	Дата последнего обновления
CRM	Robert	Smith	23/05/1988	01189 407 600	M	21/02/2013
Гарантия	Robert	Smith	23/05/1988	01189 407 600	M	05/01/2018
Центр обслуживания	Robert	Smith	23/05/1988	01189 407 600	M	13/07/2012
Маркетинг	Robert	Smith	23/05/1988	01189 407 600	M	02/05/2014
Финансы	Robert	Smith	23/05/1988	01189 407 600	M	27/05/2011

Давайте посмотрим на это приложение в действии...

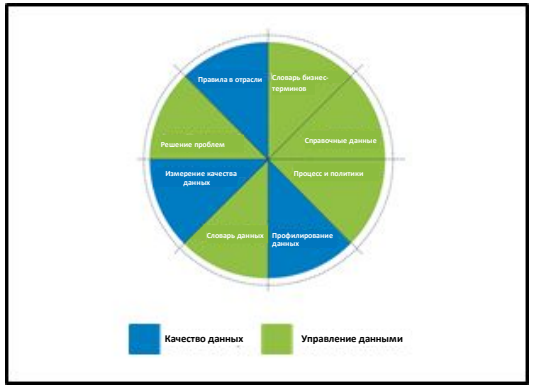
Программные решения Trillium и управление данными

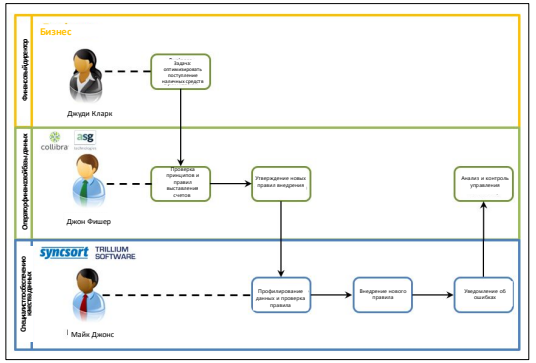
- Между обеспечением качества данных и управлением ними существует особенная взаимосвязь: имея разные задачи, эти процессы объедены единой целью — получить высококачественные данные.

- **Управление данными** предполагает разработку принципов, процессов, правил, назначение исполняющих и ответственных лиц, которые помогают организации управлять данными как активом. Оно зависит от качества данных.

- **Цель обеспечения качества данных** — сделать данные «пригодными для использования»: практического применения, принятия решений и других задач. Высококачественные данные вписываются в рамки эффективно управляемой системы правил и принципов.

- Компания Trillium Software выработала надежную схему сотрудничества и технической интеграции с двумя крупнейшими компаниями, которые занимаются управлением данными, — Collibra и ASG.

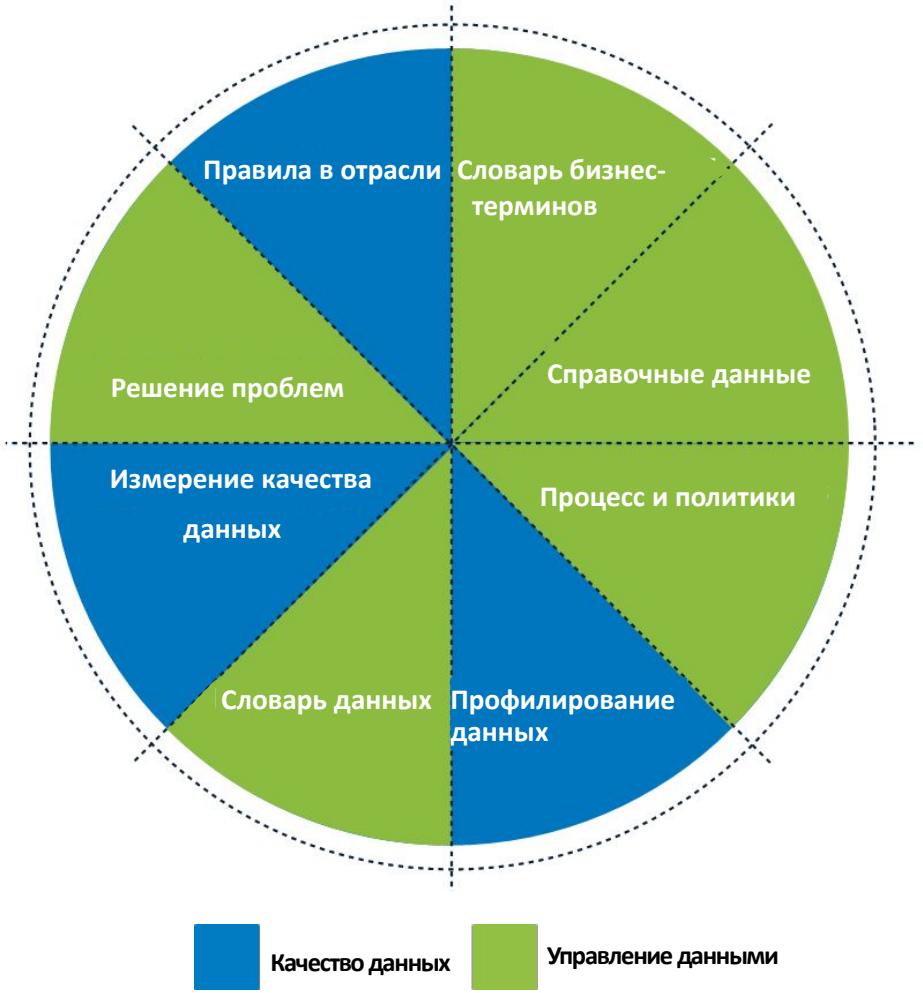


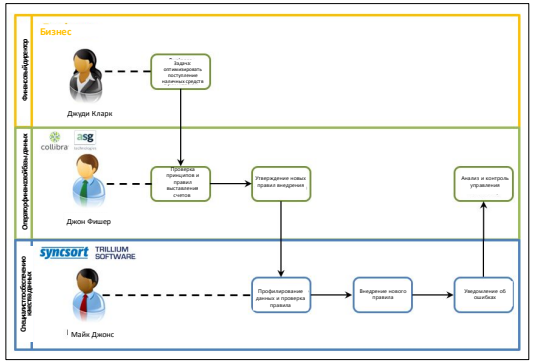


asg technologies

collibra

Программные решения Trillium и управление данными

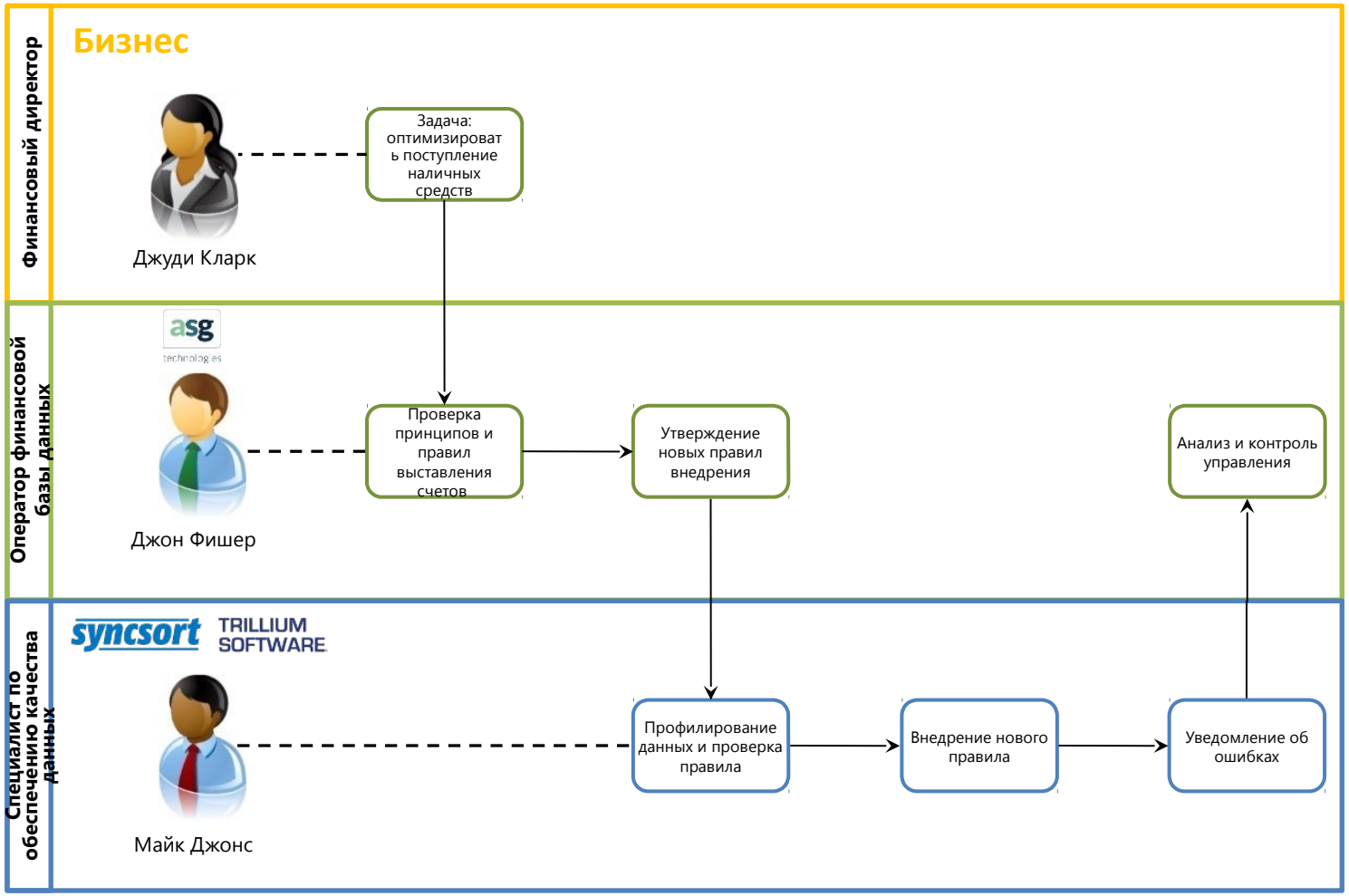


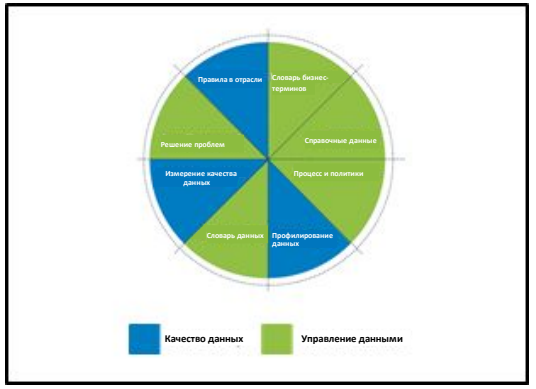


collibra

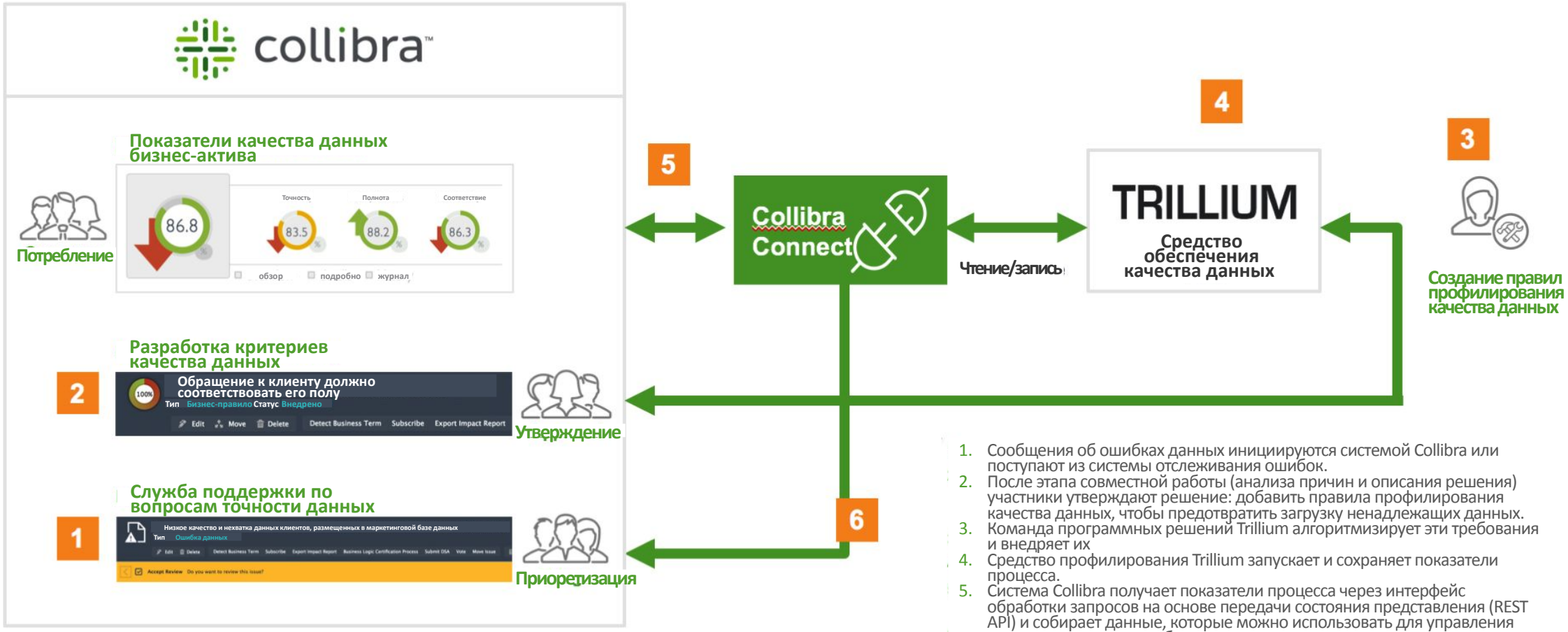
asg technologies

Пример процесса управления данными



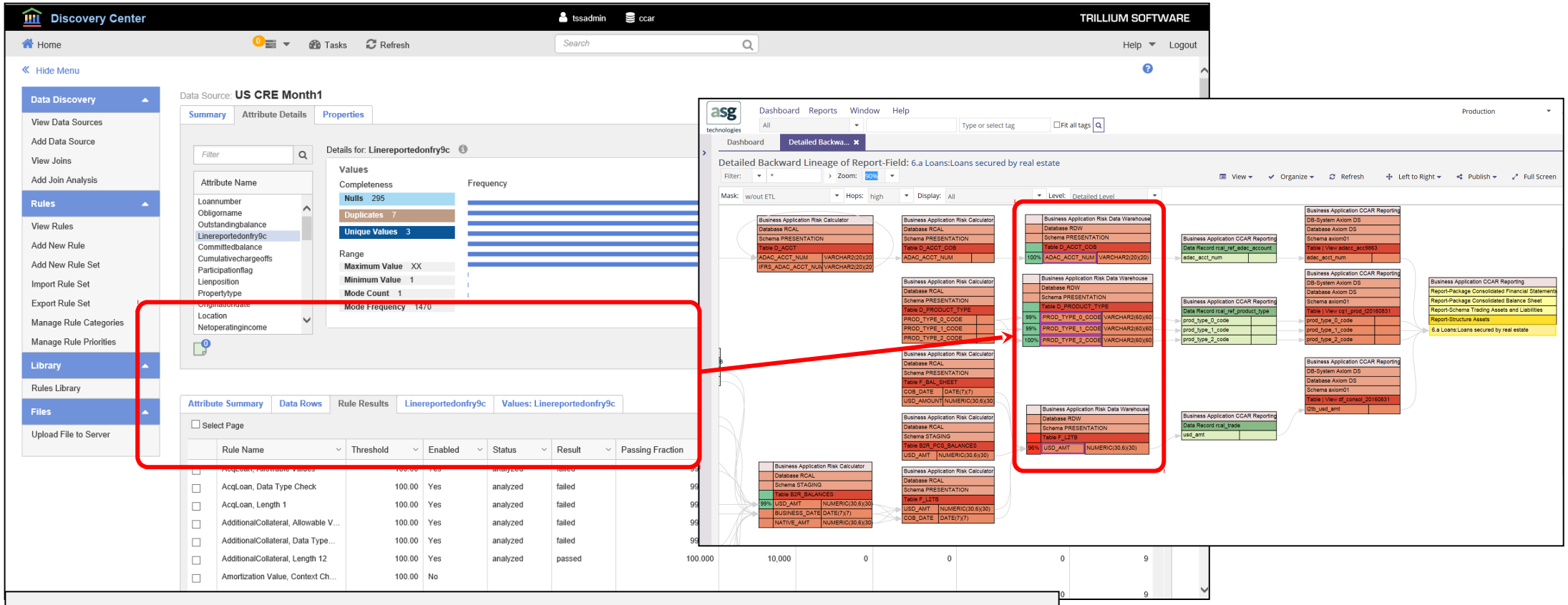


Trillium Discovery и управление данными с помощью Collibra Connect



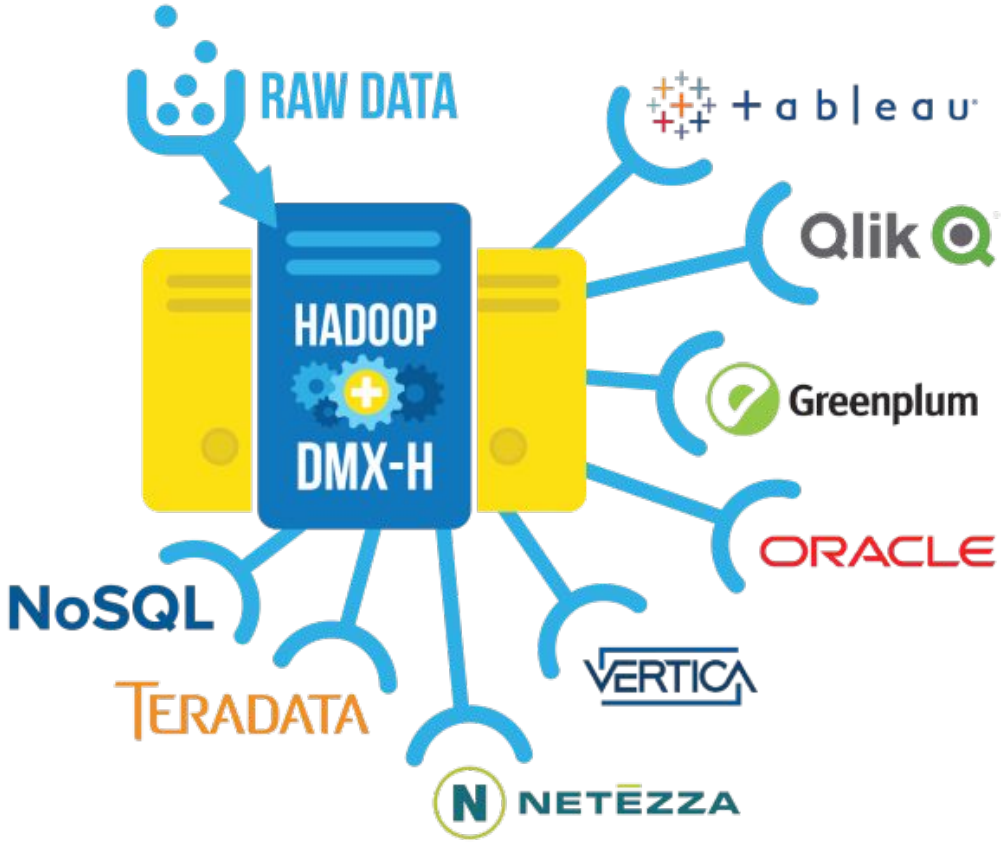
1. Сообщения об ошибках данных инициируются системой Collibra или поступают из системы отслеживания ошибок.
2. После этапа совместной работы (анализа причин и описания решения) участники утверждают решение: добавить правила профилирования качества данных, чтобы предотвратить загрузку ненадлежащих данных.
3. Команда программных решений Trillium алгоритмизирует эти требования и внедряет их
4. Средство профилирования Trillium запускает и сохраняет показатели процесса.
5. Система Collibra получает показатели процесса через интерфейс обработки запросов на основе передачи состояния представления (REST API) и собирает данные, которые можно использовать для управления данными и связывать с бизнес-терминами.
6. При превышении установленных порогов сообщения об ошибках данных загружаются из средства обеспечения качества данных Trillium в систему Collibra .

Trillium Discovery и ASG Enterprise Data Intelligence



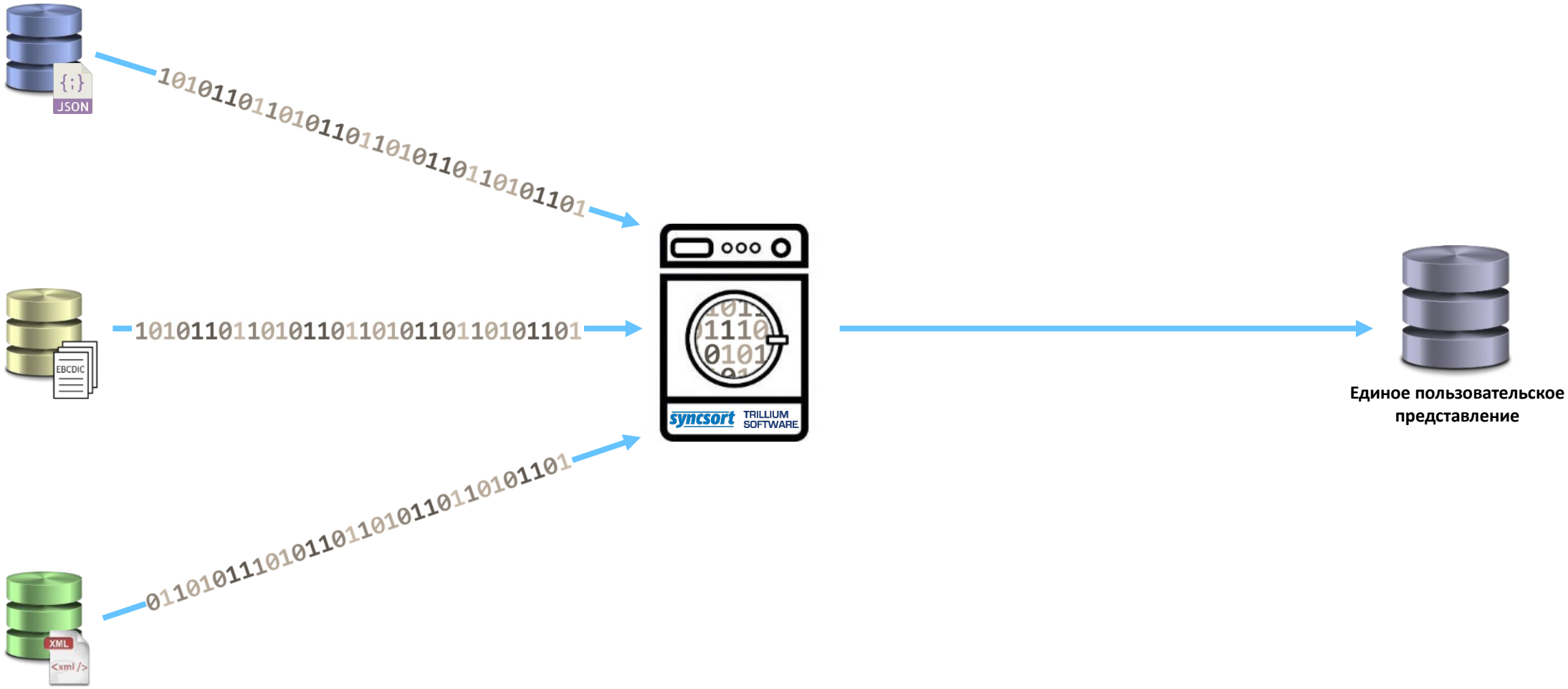
Trillium Discovery содержит **более 75 автоматизированных стандартных результатов профилирования данных,** доступных в ASG EDI.

DMX и DMX-h

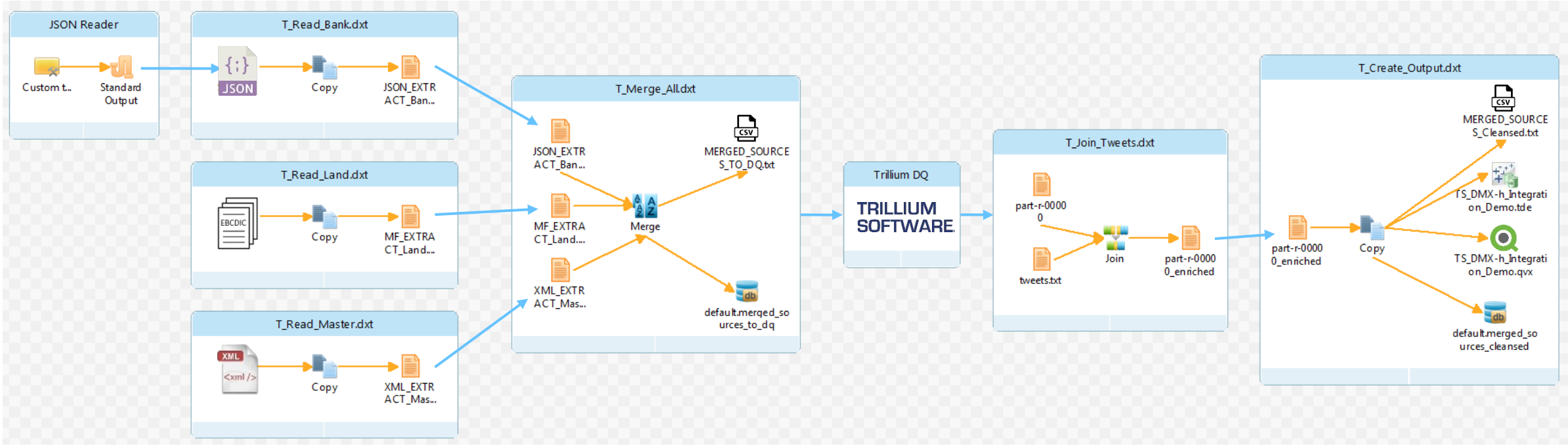


- Средство извлечения, преобразования и загрузки (ETL).
- Сбор данных практически из любого источника и их передача практически в любой источник.
- Источники пакетной обработки и потоковой передачи.
- Доступ, переформатирование и загрузка данных непосредственно в целевые источники.
- Одновременная загрузка данных из сотен таблиц в единое хранилище данных, все схемы базы данных в едином представлении.
- Более быстрая загрузка большего количества данных в среду Hadoop.

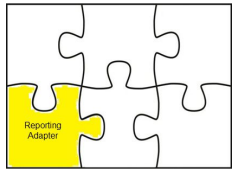
DMX позволяет превратить описанную ранее теоретическую модель...



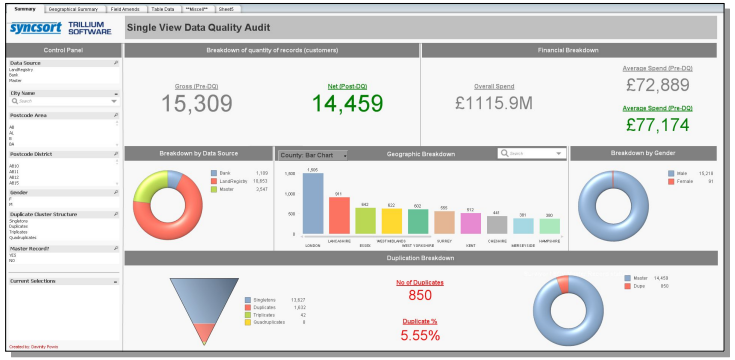
…в налаженный технический процесс DMX



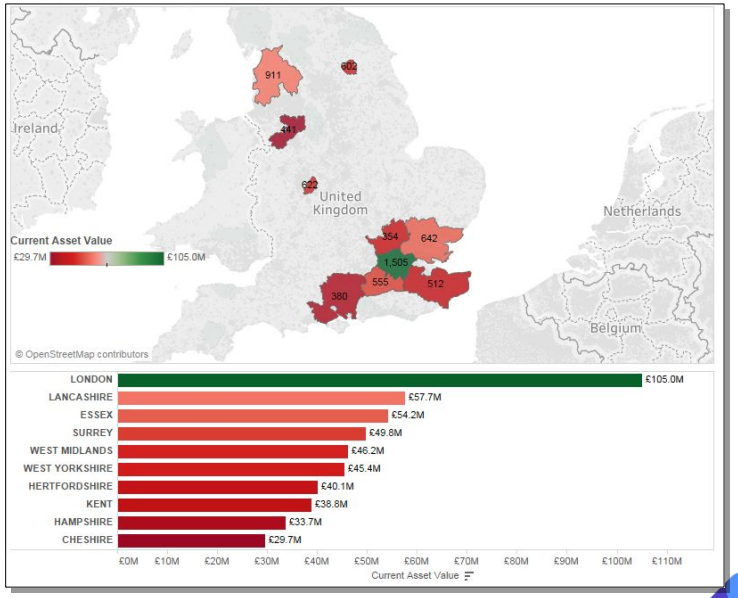
Reporting Adapter



- Компания Trillium наладила расширенную сеть взаимодействия со сторонними поставщиками.
- Это позволяет пользователям подключать средства обеспечения качества данных Trillium к инструментам бизнес-аналитики и панелям мониторинга сторонних поставщиков (например, Excel, QlikView, Tableau и т. д.).
- Это позволяет просматривать результаты, аналитические отчеты и метаданные в любом пакете программного обеспечения и среде, которую использует клиент.

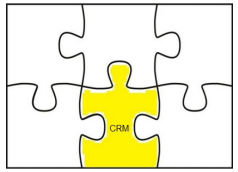


QlikView



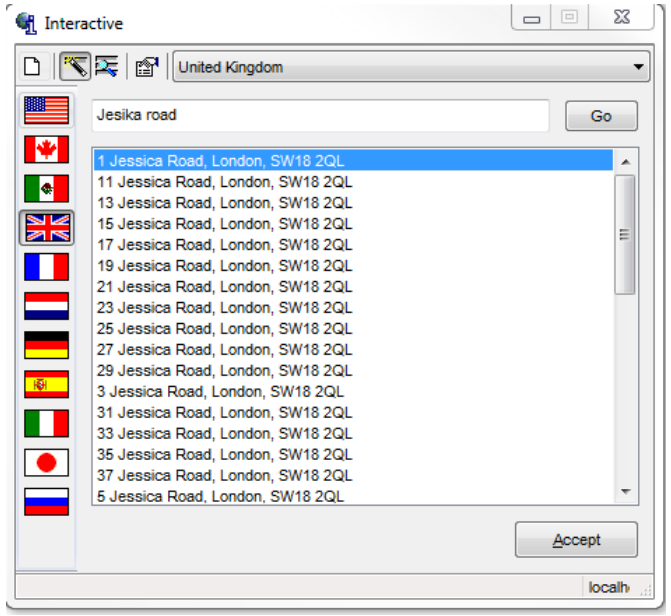
tableau

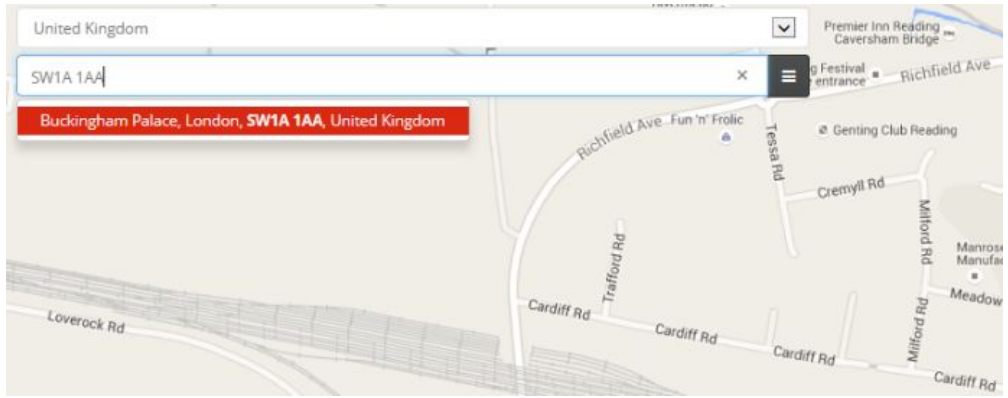
Trillium для CRM



- Компания Trillium наладила расширенную сеть взаимодействия со сторонними поставщиками.
- Комплексная интеграция с помощью API с использованием сервера приложений Trillium Director позволяет выполнять функции программных решений Trillium в продуктах сторонних поставщиков.
- Мы предоставляем подключение к системам SAP и Microsoft Dynamics.
- Это позволяет клиентам использовать процессы обеспечения качества данных непосредственно в приложениях своих CRM-систем, а также выполнять внешнюю пакетную очистку.
- Таким образом, клиенты могут управлять качеством данных, поступающих в их базу, **до** их сохранения, сокращая рабочую нагрузку каждого последующего цикла пакетной обработки.

Global Locator (GL)





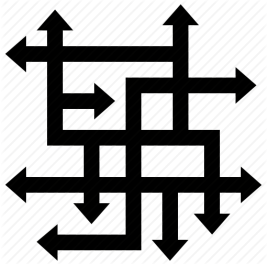
- GL — это приложение, которое быстро распознает точный адрес любой точки взаимодействия (например, в точках розничной торговли, CRM-системах, центрах обработки звонков и в сети) в интерактивном режиме.
- Стоит начать вводить адрес, и программа автоматически предложит подходящие варианты.
- Это не только обеспечивает точность и последовательность ввода данных, а также гарантирует использование правильного поля.
- Сокращение до 90 % объема вводимых пользователем данных!
- Доступно как пакетное приложение или решение, работающее в реальном времени
- Доступно в 238 странах и территориях
- В отличие от Trillium Quality, это приложение помогает управлять адресами, а не именами.

Краткие выводы

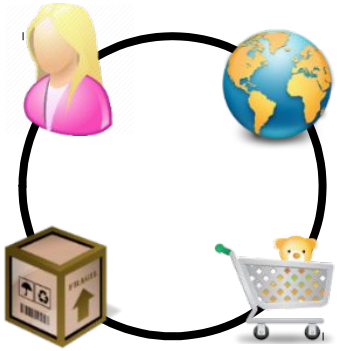
5 ключевых фактов

Лидер рынка

Сложная и масштабная архитектура



Совместимость со всеми типами данных



Широкая представленность и высокая производительность

Данные очень важны,

…но еще важнее содержащиеся в них <u>сведения,</u> которые помогают принимать более эффективные решения в бизнесе